DOCUMENT RESUME

ED 151 418 TM 006 954

AUTHOR Crambert, Albert C.
TITLE Estimation of Validity for Criterion-Referenced Tests.
PUB DATE Apr 77
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Analysis; *Criterion Referenced Tests; *Item Analysis; Norm Referenced Tests; *Test Construction; Test Items; *Test Validity

ABSTRACT

It is usually assumed that the procedures followed in developing criterion-referenced (CR) educational achievement tests ensure their content validity. This assumption is reasonable for the class of "domain-specified" CR tests, in which the item domain is completely specified by item generation rules. However, most CR tests are of the "content-specified" type, which are developed from more general content specifications which describe but do not determine the item domain. Similar considerations of validity apply for content-specified CR tests as for norm-referenced (NR) tests which are developed in a similar manner. Content validity is even more important for CR tests than for NR tests, because there is no normative group and seldom is there other evidence of the validity of the test. The CR test developer attempts to develop a test which represents the set of criterion performances which the test purports to measure, but it is actually very difficult to determine how well this goal has been achieved. The best assurance of CR test validity is careful adherence to high standards for test development. (Author/MV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ALBERT C. CRAMBERT

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED151418

ESTIMATION OF VALIDITY FOR CRITERION-REFERENCED TESTS[1]

Albert C. Crambert

U.S. Office of Education, Philadelphia, Pa.[2]

Validity is the major property of an educational achievement test. In a general sense, validity is the extent to which a measuring instrument measures "what it ought to measure." Compared to the extensive literature on techniques to determine reliability, the validity of educational achievement tests hardly is dealt with in the literature.

Of the three types of test validity defined in the Standards for Educational and Psychological Tests (American Psychological Association, et al., 1974) (content, criterion-related, and construct), it is content validity which usually is most relevant to educational achievement tests. Achievement tests are almost always expected to possess content validity regardless of whatever other types of validity may also be applicable to them. The Standards describe content validity as follows:

> Evidence of content validity is required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent. . . . Definition of the performance domain, the users' objectives, and the method of sampling are critical

TM006 954

---

[1]Presented at American Educational Research Association, New York, April 6, 1977.

[2]This paper was written in the author's private capacity. No official support or endorsement by the U.S. Office of Education is intended or should be inferred.

> to claims of content validity. . . . Content validity is determined by a set of operations, and one evaluates content validity by the thoroughness and care with which these operations have been conducted [American Psychological Association, et al., 1974, pp. 28-29].

Content validity is often referred to as curricular validity. Content validity implies far more than face validity.

The content validity of a test arises from the procedures used in constructing it. Content validity is basically self-defining in nature. It assumes that there is an underlying universe of behaviors, even if the specification of this universe is not made explicit in the development of the test.

For educational achievement tests, the test itself is usually regarded as the best available measure of the behaviors under consideration, so that no external criterion is partiularly relevant. The determination of content validity usually rests primarily on rational, subjective, and judgmental rather than statistical grounds, rooted in the objectives and content of instruction. Cattell and Butcher (1969) have observed that "validity has come cheaply in educational [achievement] measurement, because one can usually tell that a test is valid simply by reference to a content area [p. 91]." Even more than for other types of tests, the validity of an educational achievement test depends on the user's judgment as to his purposes and needs. Subjective judgment is inherent in the process of content validity.

The considerations involved in determining the validity of a criterion-referenced (CR) test are substantially the same as for norm-referenced (NR) achievement tests, with the possible difference that the validity of CR tests is usually viewed in

very practical terms, in its usefulness in making instructional decisions. Validity for a CR test can be considered to be its ability to make the appropriate decision: making neither a false positive nor a false negative decision error in determining whether an objective has been met (Meredith & Sabers, 1972). Because that instructional decision becomes self-fulfilling, it is especially crucial that a CR test used this way possess content validity in relation to the objectives of instruction (Kriewall, 1972).

It generally seems to be assumed that the procedure followed in developing a CR test ensures its content validity so that not much specific attention needs to be given to the issue of CR test validity. However, this view is not universally held (e.g., Messick, 1974). Even though there is no comprehensive presentation of a theory of validity for CR tests, there is nevertheless a substantial body of literature addressing various aspects of validity for CR tests which points out the theoretical and practical problems involved.

It is helpful to consider a distinction between two types of CR tests. One type may be referred to as "domain-specific" CR tests, and the second type as "content specified" CR tests. The considerations in validity for the two types are somewhat different.

The prototype of a domain specific approach is Hively's domain referenced testing (Hively, 1974). Other CR approaches which also fit into the domain specific category include item sampling, Bormuth's (1970) approach, and Ebel's (1962) "content standard" approach. Although there are some important differ-

ences among them, what these approaches have in common is that a domain of behaviors of interest, as reflected in test items, is completely specified in some manner. The issue of "universe specification" is crucial for domain-specific CR approaches (Cronbach, 1971).

The complete pool of items making up the domain or universe may or may not actually be constructed, but the parameters of that universe are unequivocally defined. There is no room for doubt as to whether a particular test item is or is not included in the domain. The items which make up a test form are sampled from that domain in some random or systematic manner. The key to the process of constructing domain-specific tests is the establishment of item generation rules, such as the item forms used by Hively or some other sort of very specific rules governing item construction.

Because of the close control over test content which is inherent in the procedures for developing domain-specific tests, the question of their validity is somewhat different from the consideration of validity of NR or CR achievement tests developed by more common approaches such as item construction based on a content-by-process grid. In effect, the validity question for domain-specific CR tests resolves to a matter of sampling: to what extent are the actual items representative of the domain of all possible items admissible under the item generation rules?

A number of writers (e.g., Baker, 1973; Randall, 1972) have concluded that content validity for domain-specific CR tests follows automatically from the rationale establishing

the test domain, and that no further specific consideration of their validity is necessary. From this point of view, reliability (the stability of results across samples of items) is considered to be a sufficient, rather than just a necessary, condition for validity (Jackson, 1970).

A strict operational view of the validity of domain-specific CR tests would argue that what the test measures is defined by the item generating rules. A domain-specific CR test may be construed to be a strict application of the principles of construct validity (Haladyna, Note 1). However, a too-narrow view of validity would limit the usefulness of a domain-specific test. Conventional (criterion-related) validity studies can also be carried out for domain-specific tests in order to broaden the interpretations made from them.

Because of the self-defining nature of domain-specific CR tests, the scope of possible empirical research on the question of their validity is limited. However, there have been a few studies which help to confirm the high content validity of such tests (e.g., Kwansa, 1972). The content validity of domain-specific tests within the domain defined by the test specifications seems certain to be high and readily demonstrable. However, questions concerning the validity of domain-specific CR tests for broader educational purposes must be answered by other types of validity studies.

It appears that the domain-specific CR testing approaches may be looked at as particularly rigorous applications of general content validity principles requiring specification of test content. In those situations where they can be applied,

it seems likely that domain-specific approaches could offer the advantages of precision of measurement and interpretability which are claimed by their proponents. However, requirements for their development are so demanding and the situations where they can be practically applied are quite limited (Popham, 1975), even though they could, in principle, be more widely used. A domain which can be specified to the degree of precision required for one of the domain-specific approaches is likely to be of only limited educational significance. Despite their conceptual purity, the domain-specific approaches do not appear to provide any widely generalizable solution to the issue of defining validity for CR tests.

The second type of CR tests may be referred to as "content specified" CR tests. These are CR tests developed by a procedure quite closely analogous to the traditional procedure for developing a NR achievement test (e.g., Kosecoff & Fink, Note 2). By far the majority of CR tests are developed by a "content-specified" approach, on the basis of a listing of intended outcomes of instruction, a table of specifications, or other means of circumscribing the intended content of the test. Item analysis usually is involved. In content-specified CR tests, the universe of items can be described but not fully defined, and the particular test is considered to be only illustrative of the universe rather than a true sample of it.

There are two principal differences between the process for developing a content-specified CR achievement test and a conventional NR achievement test: (a) no normative group is

involved for the CR test, so the validity of the test must rest solely on its content; and (b) because of the importance of content per se in giving meaning to performance on the CR test, there is great emphasis on the strength of the relationship which can be inferred between test content and the objectives or specifications underlying the test.

The question of content validity for the content-specified class of CR-type tests can be stated as: "How well does the content of the test represent the objectives or the specifications for the test?" This judgment can be made more precisely if the objectives have been carefully developed and are stated behaviorally. For the general case of content-specified CR tests, these objectives or test specifications are intended to describe a set of criterion performances or behaviors (the "criterion" to which the CR test is referenced), and CR measures are validated in terms of the adequacy with which they are judged to represent the criterion.

The apparently simple question of how well a content-specified test represents the set of criterion performances is actually very difficult to answer. There are no accepted quantitative methods for the task, so the question ordinarily must be answered by a process of refined judgment. It cannot be taken for granted that content-specified CR tests will possess a sufficient degree of content validity, and a rigorous validation procedure is needed in order to ensure a desirable degree of validity.

How to determine or ensure a close relationship between test objectives and test content has historically been a

major source of difficulty in achievement test construction, since long before the current discussion of CR tests. There is no infallible procedure to ensure that test items will exactly reflect the desired performance without the introduction of such extraneous, unmeasurable, and potentially potent variables as differences in perceptions between the item writer and examinee, "test-wiseness," reading ability, unobservable mental processes, physical constraints on performance, or the unintended biasing of test content through item selection.

Despite careful application of the practice of constructing achievement tests according to a table of specifications, one may question the degree to which test performance actually represents the stated objectives underlying the test. While a content-by-process grid is useful in broadening the representativeness of an achievement test, particularly in encouraging the use of items involving "higher" mental processes, such a grid does not specify a universe of possible items.

Performance objectives imply domains of test content which are considerably less open to differing interpretations than the test domains inferred from the usual table of specifications. However, performance objectives do not lead unequivocally to performance domains (Skager, Note 3), and the analysis of complex domains is still arbitrary even if the agreement of experts is achieved. Even a highly specific objective could have a potential item pool of thousands of items of which only a relative few will actually be tested

(Klein & Kosecoff, 1973), so that there is the potential for a wide gap between the objectives of testing and actual test content.

The basic guidance on the use of content specifications in developing achievement tests applies to content-specified CR tests as it does to NR tests. However, because of the particular significance in CR measurement of the relationship between instructional objectives and test content, it is necessary that the test development procedure be designed and executed with greater care, and with higher standards for consensus judgment than are usually thought to be necessary for NR tests. There is evidence that a rigorous process for collecting and analyzing the judgments of content specialists can achieve a high degree of consensus and provide a practical alternative to the complexities of a domain-specific approach or the limited usefulness of empirical data in determining CR validity (Rovinelli & Hambleton, 1976).

It appears that there would be only two relatively minor differences between the process of development of a content-specified CR test and a NR test, the process which is presumed to produce content validity. First, the CR test would be more likely than a NR test to be unidimensional or homogeneous, so the form of the specifications may be somewhat different. The objectives governing the instructional situation may also clearly imply the test content, such as in the IPI system. Second, less emphasis may be placed on item analysis of the CR test. Indeed, item analysis might not be carried out at all.

Therefore, even greater care would need to be devoted to the process of item construction. These are more differences in degree of emphasis on various aspects of the process than significant differences in the process, however.

In the more restrictive case of domain-specific CR tests, these two differences are magnified. Domain-specific tests are usually homogeneous by intent. If the test as a whole is not homogeneous, then it is homogeneous within each of a number of subtests composing the overall test. Items are generated according to strict specifications which require relatively little judgment in the actual item construction. Item analysis is rarely conducted for a domain-specific test; the principles of the various domain-specific approaches provide that the item generation rules are a sufficient guide to the production of items, and any further selection of items on the basis of their statistical properties would be a violation of the item generation rules.

While the validity of a content-specified CR test is best dealt with at the outset through the process of test construction, there have been suggestions for indirect approaches to inferring or confirming content validity after the fact. CR tests could be validated through the interpretation of item data from pretest and posttest, provided that the item is sensitive to instruction (Lindvall & Cox, 1970; Wilson, 1973). Such an interpretation would be made <u>after</u> the test items have already been shown through item analysis to be reliable and valid for the objective being measured. A different use of

item data to establish content validity for CR tests was suggested by Klein and Kosecoff (1973), who noted that an item should correlate more highly with other items intended to measure the same objective than it does with items measuring other objectives.

To what extent are statistical indices of validity useful for CR tests? Although content validity is usually the most important type of evidence of validity for educational achievement tests, achievement tests can also be used in a criterion-related sense, usually to predict future status on some variable. It may be convincingly argued that all achievement testing carries the implication of prediction of some future measure. Knowledge of present status alone is of limited value unless it can also be used to infer future performance: the ability to succeed in future instruction, the likelihood of equivalent standing on some future measure, etc. When an attempt is made to specifically relate achievement test performance to some other variable (usually prediction of a future variable), the paradigm of criterion-related validity applies.

The major consideration in the use of the conventional statistical indices of criterion-related validity with CR tests would arise from the possibility of reduced variance and consequently reduced reliability for CR tests when these are computed using the conventional equations. To the extent that its variance is reduced, the CR test would also have restricted correlations with external criterion variables

with which it might be correlated. The predictive validities of CR tests might therefore typically be lower than the predictive validities of NR tests against the same criterion variables.

An abstract comparison of the criterion-related validities of a CR test and a NR test against the same criterion could be made by correcting the obtained correlation between the CR test and the criterion for the restriction in CR test variance. While this correction would permit a comparison of the merits of the CR and NR tests without the complication of differing score variance, such a comparison might be unrealistic in terms of the actual use of the CR test.

While it might appear that statistical indices of validity would not be very useful for CR tests, there are three offsetting considerations:

(a) Because of the purposes and circumstances in which CR tests are usually used, their validity against an external criterion is less likely to be a matter of interest.

(b) There are many factors which mitigate against examinees achieving perfect or near-perfect scores on a CR test as they might be expected to do according to the rationale of the test. Thus, score variance on a CR test, and the test's ability to correlate with an external criterion, usually will be higher than might be expected. Accumulating evidence, both logical and empirical, suggests that the actual variance of CR tests differs little from NR tests (Haladyna, Note 1); and

(c) The validity coefficient expresses the relationship

between two variables over their entire range. However, depending upon the type of decision to be made on the basis of the validity data, which may involve only a limited part of the score range, the CR test may still be nearly as efficient in making specific decisions as a NR test which has a higher validity coefficient.

There is also a type of statistical indicator which is unique to CR tests and does not fit into the conventional framework for examination of validity: sensitivity to instruction. Given the conceptual foundation of CR testing, it is logical to include in a CR test only items which reflect gains in performance attributable to instruction. A simple index such as the Cox-Vargas Pretest-Posttest Difference Index (Cox & Vargas, 1966) will identify items on which performance is influenced by instruction. A test composed of such items would have validity in this narrow sense, although test performance would not be readily explainable in any broader context such as content validity.

It can be concluded that, with the availability of easily applied procedures for helping to establish the validity of CR tests, especially content-specified CR tests, that there can be little justification for assuming that a casual development process will produce a valid test. Careful specification and control of the process is the best guarantee of content validity, but many other techniques are available to help bolster the evidence of validity and understanding of what is being measured by the test.

Reference Notes

1. Haladyna, T. The paradox of criterion-referenced measurement. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1976.

2. Koseccff, J., & Fink, A. The feasibility of using criterion-referenced tests for large-scale evaluations. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

3. Skager, R. W. Generating criterion-referenced tests from objectives - based assessment systems: Unsolved problems in test development, assembly, and interpretation. Paper presented at the meeting of the American Educational Research Association, New Orleans, February 1973.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D. C.: American Psychological Association, 1974.

Baker, E. L. The technology of instructional development. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally, 1973.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Cattell, R. B., & Butcher, H. J. The prediction of achievement and creativity. Indianapolis: Bobbs-Merrill, 1969.

Cox, R. C., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, February 1966. (ERIC Document Reproduction Service No. ED 010 517.)

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Hively, W. (Ed.), Domain referenced testing. Englewood Cliffs, N. J.: Educational Technology Publications, 1974.

Jackson, R. Developing criterion-referenced tests. Princeton, N. J.: Educational Testing Service, 1970. (ERIC-TM Report 1.) (ERIC Document Reproduction Service No. ED 041 052)

Klein, S. P., & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. Princeton, N. J.: Educational Testing Service, 1973. (ERIC-TM Report 26.) (ERIC Document Reproduction Service No. ED 083 284)

Kriewall, T. E. Aspects and applications of criterion-referenced tests. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1972. (ERIC Document Reproduction Service No. ED 063 333)

Kwansa, K. B. Investigation of the relative content validity of norm-referenced and domain-referenced arithmetic tests. Unpublished doctoral dissertation, University of Pittsburgh, 1972. (ERIC Document Reproduction Service No. ED 077 742)

Lindvall, C. M., & Cox, R. C. Evaluation as a tool in curriculum development: The IPI evaluation program. Chicago: Rand McNally, 1970.

Meredith, K. E., & Sabers, D. L. Using item data for evaluating criterion-reference measures with an empirical investigation of index consistency. Paper presented at the meeting of the Rocky Mountain Educational Research Association, November 1972. (ERIC Document Reproduction Service No. ED 073 125)

Messick, S. The standard problem: Meaning and values in measurement and evaluation. Princeton, N. J.: Educational Testing Service, 1974. (Research Bulletin RB-74-44.)

Popham, J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Randall, R. S. Contrasting norm-referenced and criterion-referenced measures. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1972. (ERIC Document Reproduction Service No. ED 065 593)

Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976. (ERIC Document Reproduction Service No. ED 121 845)